Enhancing Child Safety: Multimodal Approach for Video Sentiment Analysis in Online Environments (Yee Sen Tan)

Agenda

- Introduction (Motivation and Goal)
- Related Works
- Proposed Methodology
- Results & Discussion
- Conclusion, Limitations and Future Works

- Approximately 75% of young children possessing their own tablets (Kabali et al., 2015), where infant as young as one started using mobile devices (Rideout & M. B. Robb, 2017)
- This exposes them to social media recommenders, such as for your page (fyp), enhancing user's experience (Cotter et al., 2022)
  - But may expose them to disturbing content and create a loop, leading to formation of filter bubbles (Yesilida & Lewandosky, 2022)
  - It has shown to cause self-harm and damage to physical/mental health (Abi-Jaoude, 2020)

There is a need to protect these children, which can be done via machine learning

Relevance in today’s climate

- In 2024, Congress has held multiple hearings on issues related **to online child safety content**. Companies such as Meta, X (Twitter), TikTok, etc. were involved (Razi, 2024)
- E.g.: Mark Zuckerberg apologized to parents of the children who died due to causes related to social media content; TikTok CEO as questioned on content-moderation

Goal is to **identify and filter** such negative content

Assign video as either

“Positive” or **“Negative”**

Positive - (child friendly)

Negative - unsuitable for child’s consumption

Negative Sentiment will span from “Slight Negative”, “Negative” to “Strong Negative”

Assign video as either

“Positive” or **“Negative”**

Positive - (child friendly)

Negative - unsuitable for child’s consumption

Negative Sentiment will span from “Slight Negative”, “Negative” to “Strong Negative”

Not an easy task due to different modalities

Text

Textual Content

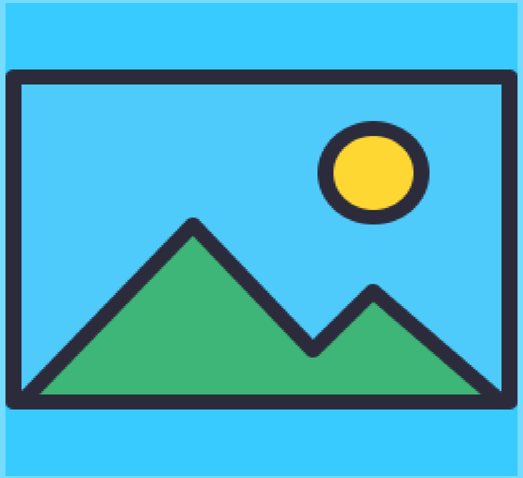
Textual information can exist in the form of speech (from audio), subtitles, text-overlays on videos

** We mainly focus on the speech for this*

Visual

Image Frames

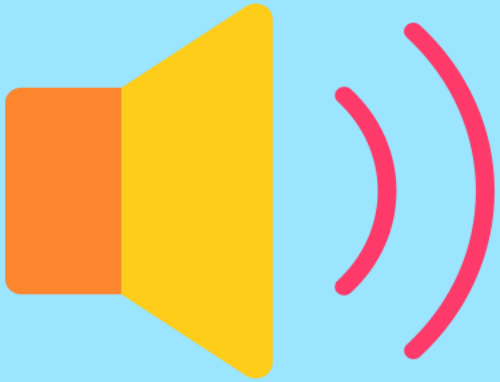
Image frames captures scenes of the videos, visual information, sequences of events, etc



Audio

Audio

Usage of background music, speeches, etc.



01

02

03

04

Multimodal Sentiment Analysis

Multimodal methodologies has advantages over analyzing a single modality. (Poria et al., 2018; Morency et al., 2011)

However, key issue can arise when different modalities interact, which could lead to disparities in individual contributions in multimodal fusion – possibly due to loss of modality-specific information, and risk of overfitting(Wu et al., 2022). There are in fact other fusion approaches, such as late-fusion that builds models for each modality and combine results via averaging or voting (Abdu et al., 2022)

Today, multimodal works includes Contrastive learning – eg: Contrastive Language Image-Pretraining (CLIP) and Bootstrapping Language-Image Pre-training (BLIP) that has shown strong downstream tasks performance (Radford et al., 2021; Li et al., 2022)

01

02

03

04

Text Sentiment Analysis

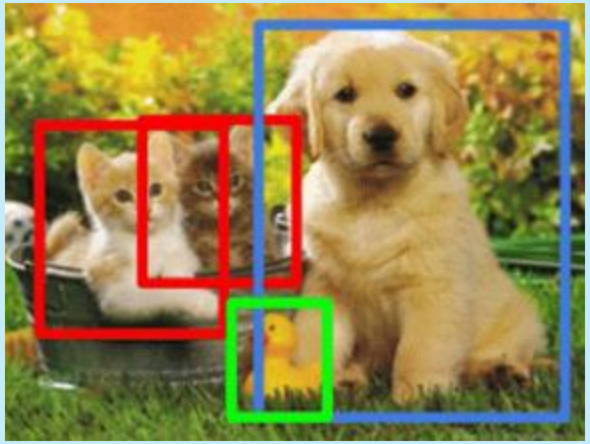
Traditional text analysis work mainly revolves around English (Nguyen et al., 2019) which may longer prove to be effective for sentiment analysis considering social media's growth, where the use of Non-English and code mixing has increased, hence the importance of multilingual text models (Ou & Li, 2020).

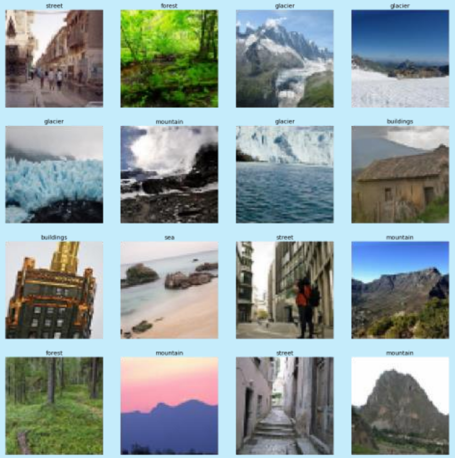
Eg: XLM-RoBERTa - transformer based pre-trained language model (Robustly Optimized BERT + Cross Lingual/XLM. Scalable and can be finetuned for downstream tasks (Pant & Dadu, 2020)

Visual Sentiment Analysis

Visual information can come in various forms

1. Objects
2. Scenes
3. Composition of style of images/videos





01

02

03

04

Visual Sentiment Analysis

Visually, elements such as surrounding scenery and actions in videos can determine a video's sentiment. For instance, violent scene that includes Gorey, Bloody, or elements of Horror are negative content that should not be displayed to children (Wang et al., 2011)

01

02

03

04

Visual Sentiment Analysis

Traditionally, most work uses Convolutional Neural Networks (CNNs) for image related tasks and includes extracting image frames before undergoing pre-processing and downstream tasks (Gunawan et al., 2020)

In recent years, Vision Transformers (ViT) is shown to outperform CNN though require more training data. However, when pre-trained with huge amounts of data, using transfer learning proves that it is superior (Dosovitskiy, 2022; Deininger et al., 2022).

01

02

03

04

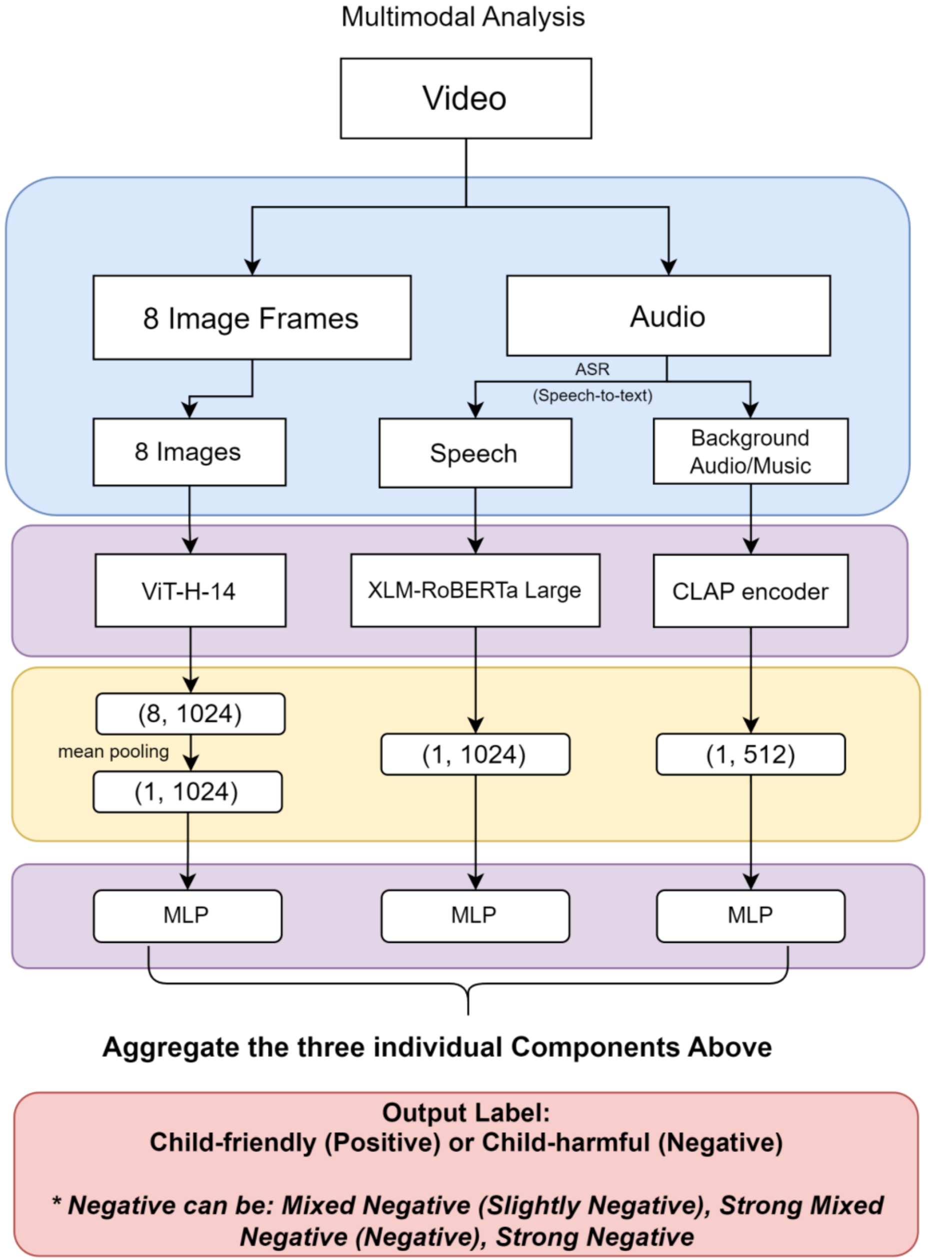
Audio Sentiment Analysis

Audio typically comes in the form of 1) speech & 2) no-speech

Speech – form of audio, mostly analyzed using Automatic Speech Recognition (ASR) to extract text, followed by applying NLP techniques.

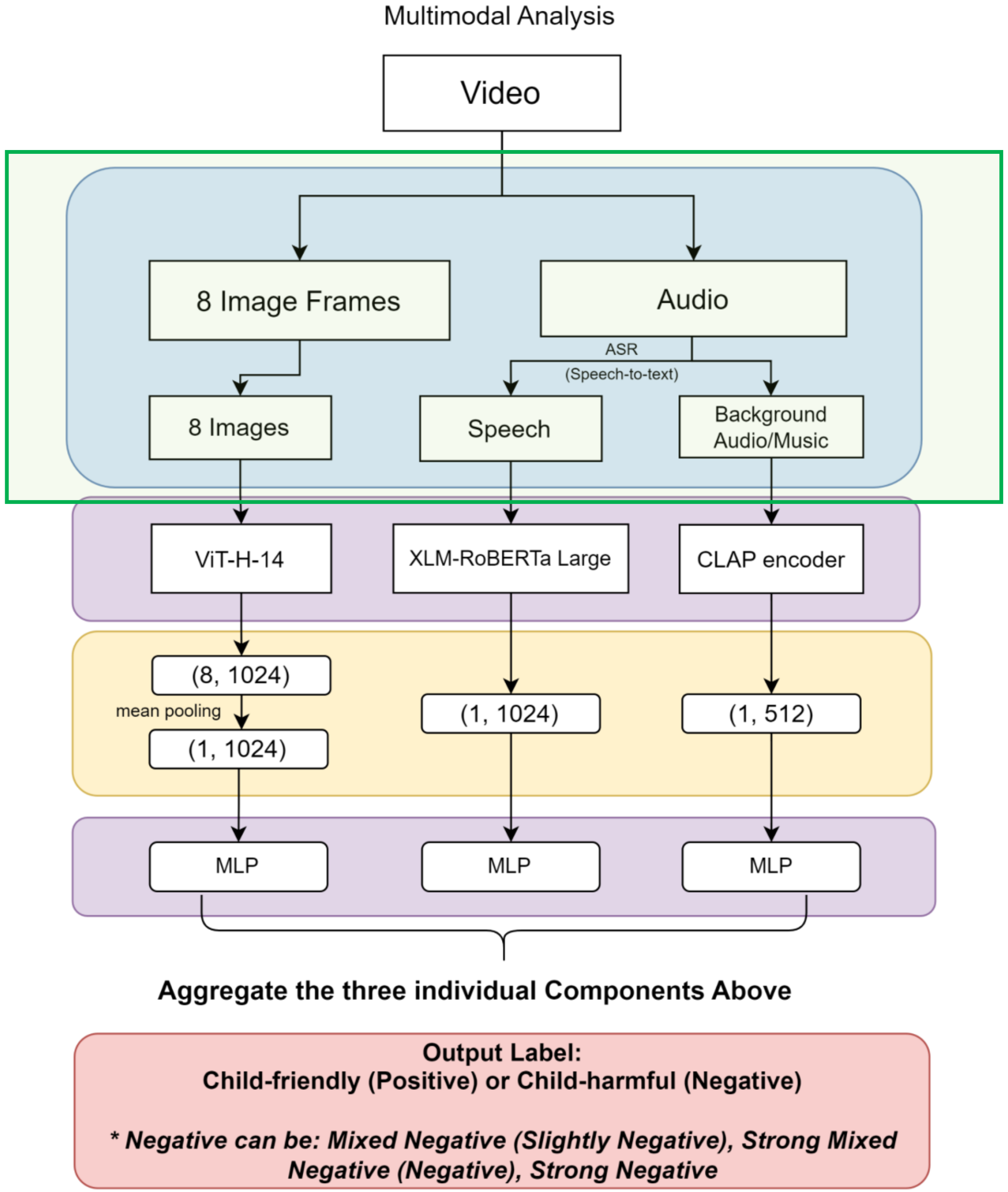
Music – is another type of audio, no-speech, that affects sentiment too. E.g.: scary background music or suspense.

Recent works in contrastive learning domain proposed Contrastive Language-Audio Pretraining (CLAP) that has also shown strong zero-shot and downstream performance for audio-related tasks (Wu et al., 2023)



PROPOSED METHODOLOGY

Late-fusion multimodal approach combining text, visual, audio. Classifier is built for each modality to determine individual modality sentiment and a proposed voting mechanism will determine the final sentiment



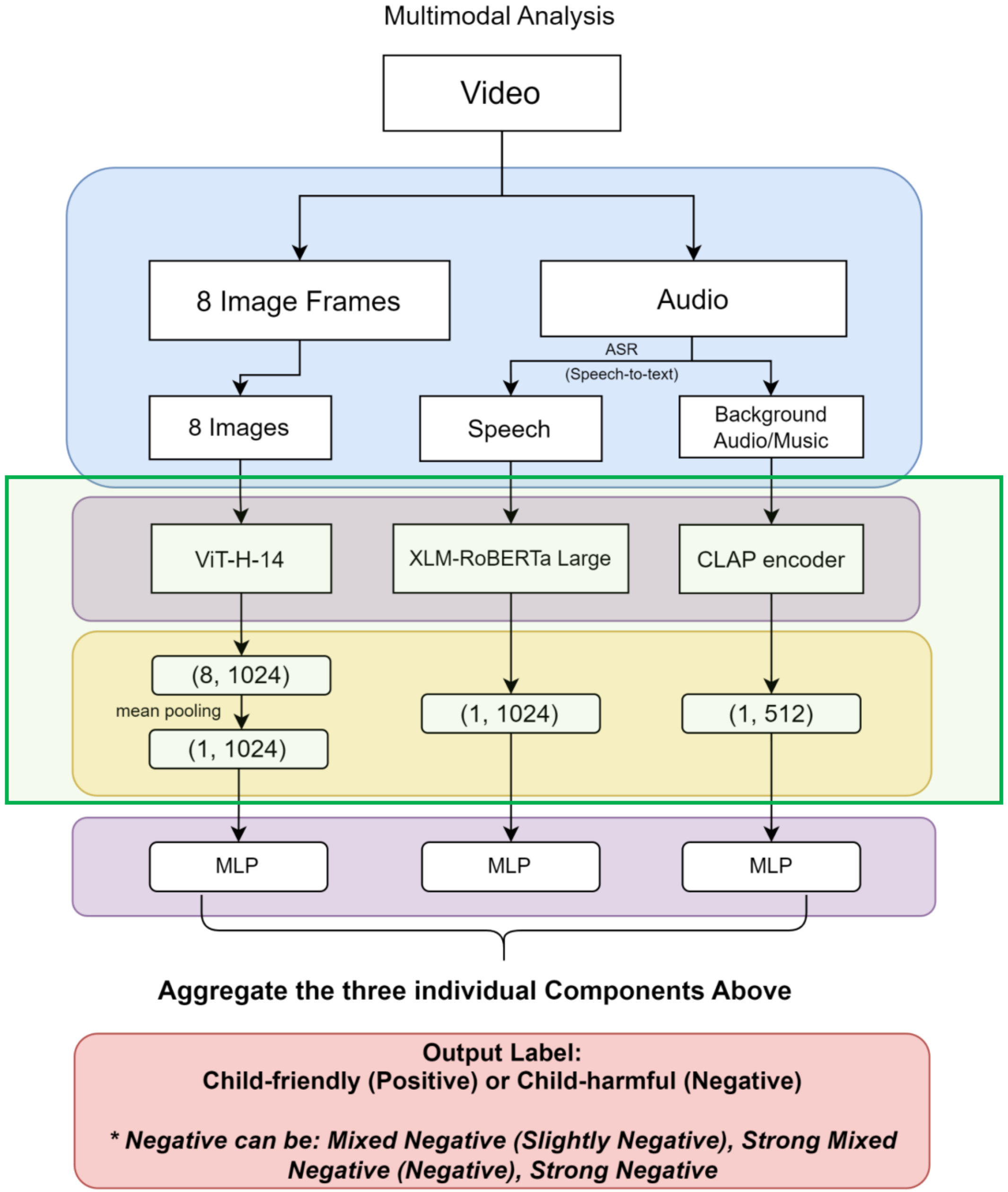
For each video

1) Extract 8 image frames Video

- OpenCV
- Equal binning approach, always returns 8 frames

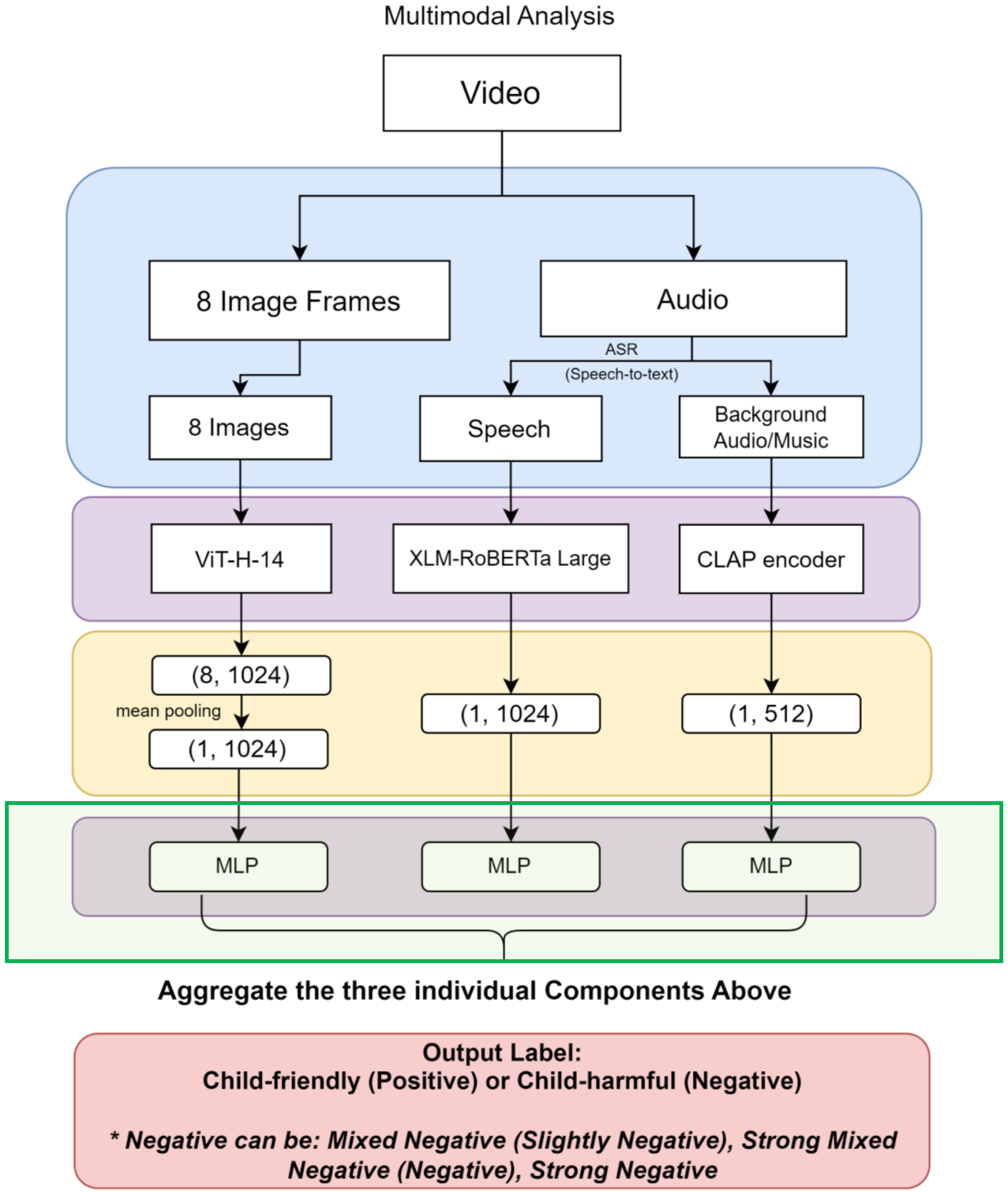
2) Extract Speech and BGM from Video

- Audio Separator Library
- OpenAI Whisper Large



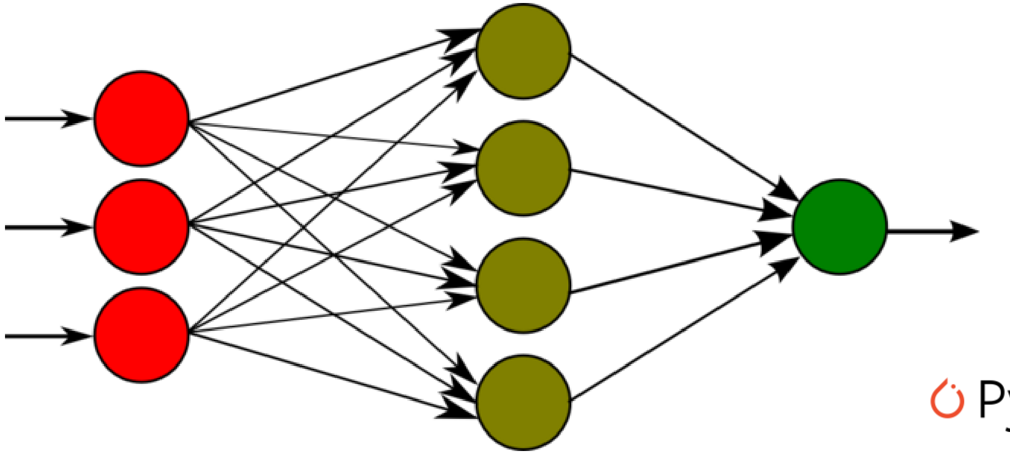
1) **laion/CLIP-ViT-H-14-laion2B-s32B-b79K**
   - VIT-H-14 (Vision)
   - XLM-RoBERTa Large (Text)
2) **CLAP General**
   - For Audio

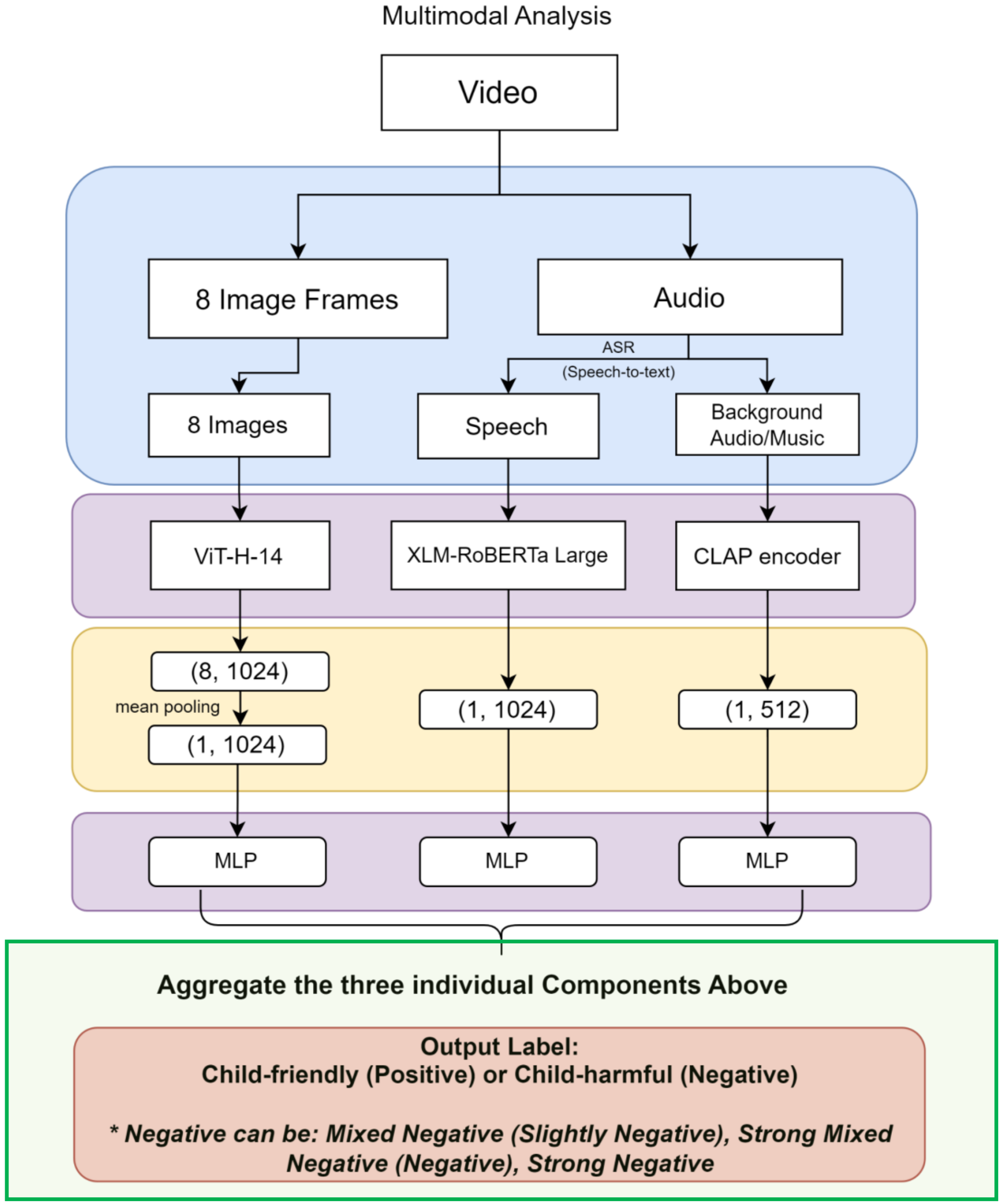
Embeddings will be obtained using these encoders



Total of 3 Individual Classifiers – training elaborated later

Multi Layer Perceptron





Each classifier will predict a sentiment (Positive/Negative)

Voting Mechanism will determine the final overall sentiment

Possible Output:

- **Positive**: All outputs positive
- **Slight Negative**: 2 outputs positive 1 output negative
- **Negative**: 1 output positive, 2 output negative
- **Strong Negative**: All outputs negative

Imagine a scenario whereby

- No horror/Gorey scene

- Audio not scary

- *Vulgarity is used (Will be detected even though it is minor)*
  Overall Sentiment: Slight negative
- *Examples shown later!!!*

Might be missed if we do early-fusion or do an "averaging"

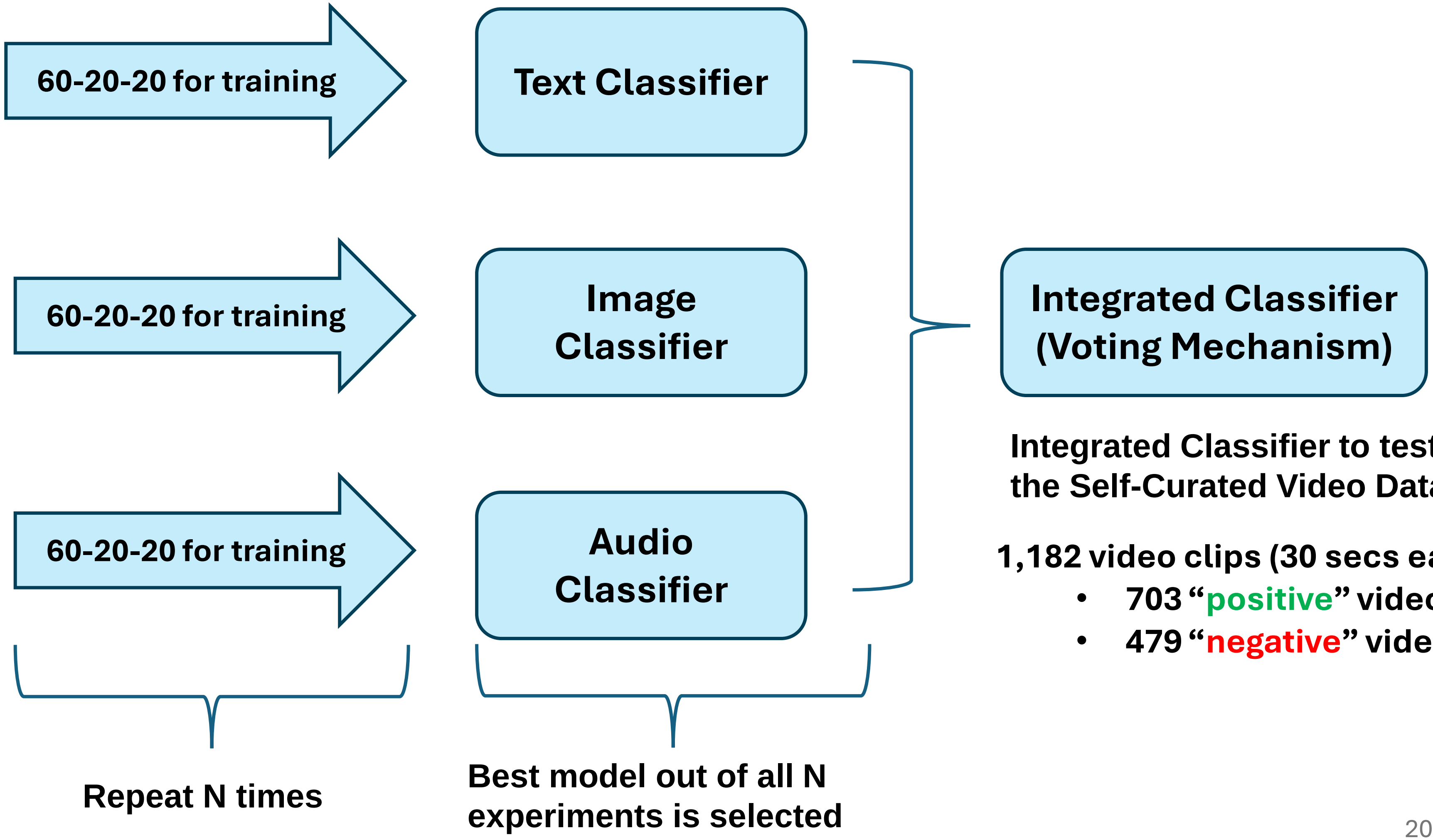
3 dataset – one for each modality

1 test dataset of videos for testing integrated classifier

Hate Speech and Offensive Language (HSOL)
2,872 positive & 7,128 negative

Self-Curated Visual Dataset
7,986 positive & 2,536 negative

Self-Curated Audio Dataset
2,173 positive & 2,080 negatives



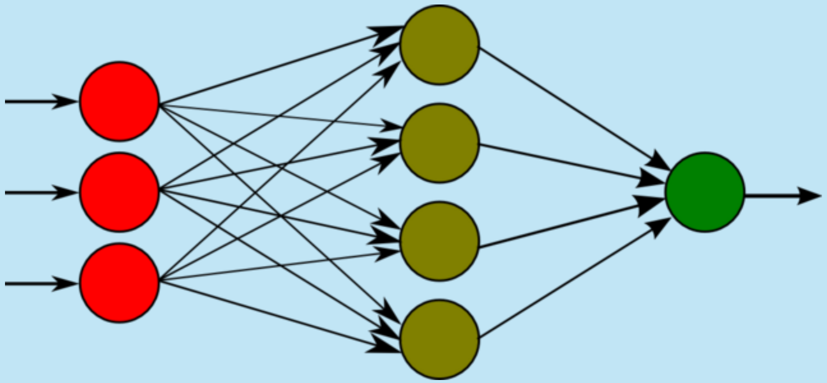
Integrated Classifier to test on the Self-Curated Video Dataset

1,182 video clips (30 secs each)
- 703 "positive" video clips
- 479 "negative" video clips

Model Training

Main Metric of Interest: F1 Score and Accuracy

- 3 Linear layer
- Dropout regularizations (30%)
- Leaky ReLU activation function
- Focal Loss (Handle imbalance)
- Sigmoid



Data used for training classifiers (Not testing video)

- 60 % Training; 20% Validation; 20% Test
- Batch Size 32
- Early Stopping 20 Epochs

- Lookahead Optimizer (Robust to parameter changes) (Zhang et al., 2019)
  - Experiment with SGD and Adam Optimizer
  - Learning Rate of 1e-2 and K=10
- Learning Rate Scheduler
  - Dynamic LR scheduler (ReduceLROnPleateau)
  - LR reduced by factor of 0.1 if no improvement >5 epochs
- Efficient Gradient Scaling with GradScaler

TABLE I
PERFORMANCE EVALUATION OF TEXT CLASSIFIER

Model	Optimizer	Loss Function	Accuracy	F1
Model 1	SGD	FL(α=1,γ=2)	93%	91%
Model 2	Adam	FL(α=1,γ=2)	94%	93%
Model 3	**SGD**	**FL(α=0.25,γ=2)**	**95%**	**94%**
Model 4	Adam	FL(α=0.25,γ=2)	71%	42%

Note that my test set for text has 1,427 negatives and 574 positive samples

Recall: Original author proposed best params to be and α = 0.25 and γ = 2

TABLE II
PERFORMANCE EVALUATION OF IMAGE CLASSIFIER

Model	Optimizer	Loss Function	Accuracy	F1
Model 1	**SGD**	**FL(α=1,γ=2)**	**96%**	**95%**
Model 2	Adam	FL(α=1,γ=2)	95%	94%
Model 3	SGD	FL(α=0.25,γ=2)	95%	94%
Model 4	Adam	FL(α=0.25,γ=2)	94%	44%

Note that my test set for image has 507 negatives and 1,597 positive samples

Recall: Original author proposed best params to be $\gamma = 2$ and $\alpha = 0.25$

TABLE III
PERFORMANCE EVALUATION OF AUDIO CLASSIFIER

Model	Optimizer	Loss Function	Accuracy	F1
Model 1	SGD	FL(α=1,γ=2)	82%	82%
Model 2	Adam	FL(α=1,γ=2)	93%	93%
Model 3	SGD	FL(α=0.25,γ=2)	82%	82%
Model 4	**Adam**	**FL(α=0.25,γ=2)**	**94%**	**94%**

Note that my test set for audio has 416 negatives and 435 positive samples

Recall: Original author proposed best params to be γ = 2 and α = 0.25

TABLE IV
PERFORMANCE EVALUATION OF INTEGRATED CLASSIFIER

Modality	Accuracy	F1	
Text + Image + Audio	**81%**	**81%**	
Text + Image	92%	92%	
Text + Audio	72%	71%	Worst
Image + Audio	85%	85%	
Text	72%	72%	
Image	97%	97%	Best
Audio	74%	73%	

TABLE IV
PERFORMANCE EVALUATION OF INTEGRATED CLASSIFIER

Modality	Accuracy	F1
Text + Image + Audio	**81%**	**81%**
Text + Image	92%	92%
Text + Audio	72%	71%
Image + Audio	85%	85%
Text	72%	72%
Image	**97%**	**97%**
Audio	74%	73%

Breakdown

Text + Image + Audio	Actual Positive	Actual Negative
Predicted Positive	482	3
Predicted Negative	221	476

Breakdown

Image Only	Actual Positive	Actual Negative
Predicted Positive	686	18
Predicted Negative	17	461

TABLE IV
PERFORMANCE EVALUATION OF INTEGRATED CLASSIFIER

Modality	Accuracy	F1
Text + Image + Audio	**81%**	**81%**
Text + Image	92%	92%
Text + Audio	72%	71%
Image + Audio	85%	85%
Text	72%	72%
Image	**97%**	**97%**
Audio	74%	73%

Breakdown

Text + Image + Audio	Actual Positive	Actual Negative
Predicted Positive	482	3
Predicted Negative	221	476

Breakdown

Image Only	Actual Positive	Actual Negative
Predicted Positive	686	18
Predicted Negative	17	461

Incorrect predictions here were due to negativity being present in the form of another modality. E.g.: audio
Total of 18 mistakes as compared to Text+Image+Audio which only had 3 mistakes
Why not text or audio alone? **Performance bad. Of course, unable to detect jump scares, gorey scenes, etc**

Later, I will show you why all modalities must work together.

TABLE V
DEEPER ANALYSIS ON RESULTS OF VOTING MECHANISM

Modality	Accuracy	F1
Text + Image + Audio	**81%**	**81%**
Text	72%	72%
Image	97%	97%
Audio	74%	73%

Actual Label	Predicted Label	Count (%)
Negative	**Negative**	**165 (35.45%)**
	Slight Negative	**131 (27.35%)**
	Strong Negative	**180 (37.58%)**
	Positive	**3 (0.62%)**
Positive	**Positive**	**482 (68.56%)**
	Slight Negative	**201 (28.59%)**
	Negative	**20 (2.85%)**

Image alone perform the best, followed by our integrated classifier, then the audio and text

However, the false predictions are extremely mild.
For instance, false positive only accounts for 0.62%.
Even though there is **almost 30% False Negatives, most of them are "Slight" Negative**, which is negligible)

One Quick Example of Image alone being insufficient in capturing negativity

No jump scare, don't worry!

Would you rather

(1) Filter some child-friendly videos away with almost **NO harmful video left**

OR

(2) More videos but **HIGHER risk of harmful content** (vulgar, gory, horror, etc)

Even though my methodology **drops F1 score (more False Negatives)**, the methodology **should still be adopted**

Model Explainability

Local
Interpretable
Model-agnostic
Explanations

LIME (Ribeiro et al., 2016)

How it works?

1.Generating Perturbations: LIME generates perturbations, by randomly masking or modifying data features

2. Model Prediction: LIME then feeds these perturbed samples through the model to obtain their predictions.

3. Explain: Capture how swapping out each data/feature changes the prediction and identify the importance/contributor

Model Explainability

Investigating Text-Related performances

Video Time

Cartoon - “Family Guy”

Model Explainability

Blue represents Negative

Text with highlighted words

Well, I guess that means I can get rid of all my grandma merch. I'll just donate it to Goodwill. You know what gilf means, right?

*** Note how accurate the ASR is, and how it manage to pick up the negative words**

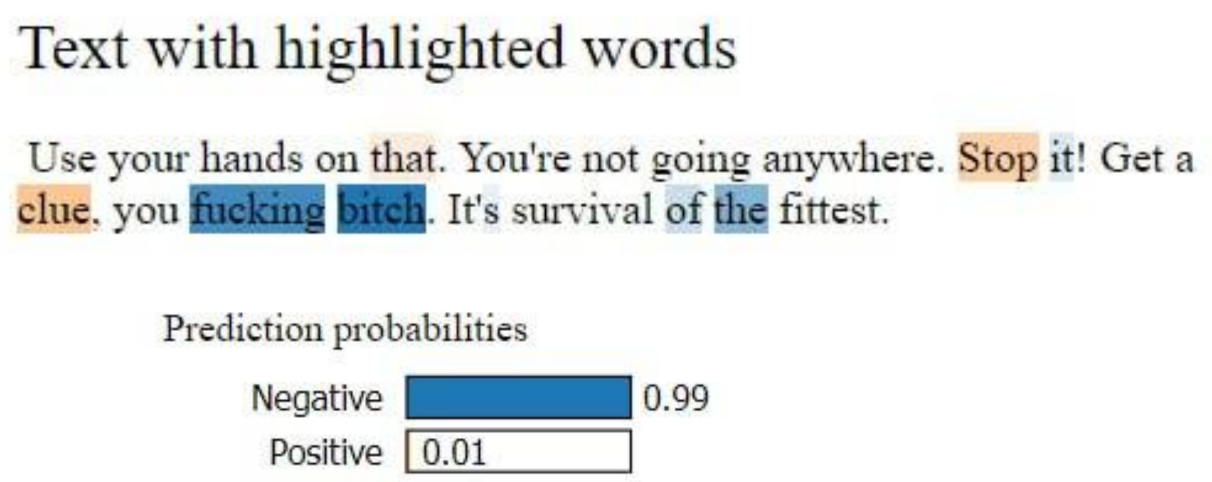
What does GILF mean?

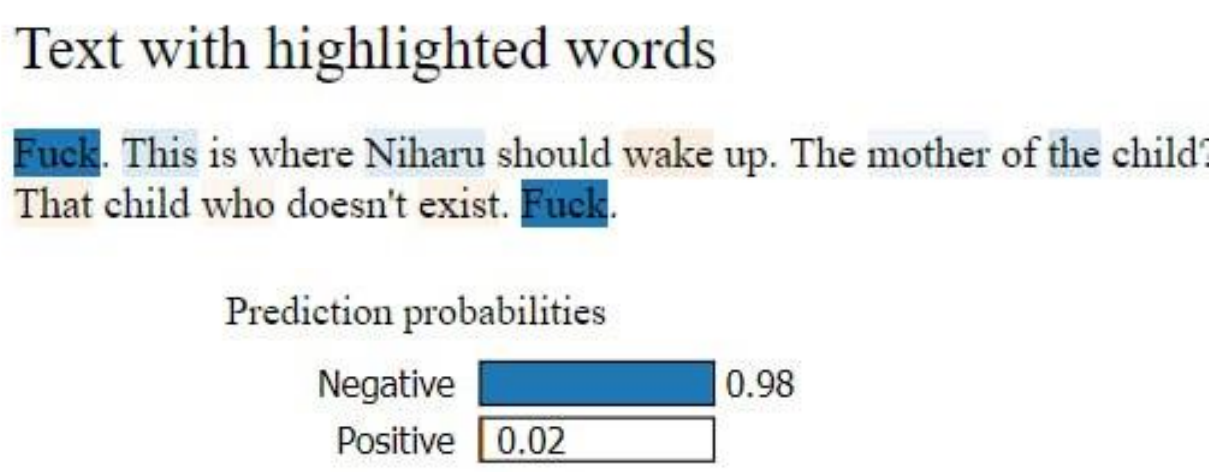
12 Nov 2018 — GILF is an acronym that stands for "**Grandmother I'd Like to Fuck**." It is an offensive and disrespectful slang term that objectifies or ...

Model Explainability

Investigating Text-Related performances

Many more examples





Model Explainability

Investigating Image-Related performances

Examples

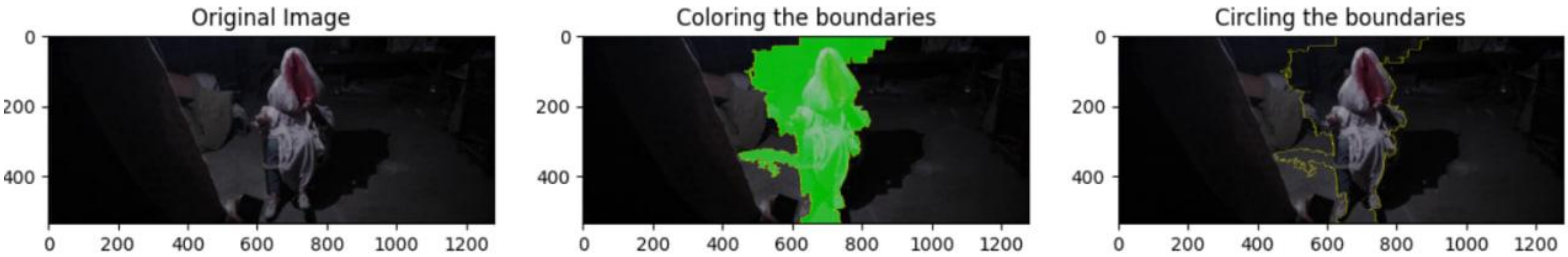
Horror Movie – The Conjuring 2

Model Explainability

Investigating Image-Related performances

Examples

Horror Movie – Insidious



Summary on WHY multimodality and voting is still preferred

1. **Instances whereby single modality cannot capture stuff**
    - Recall the video of the women running in a haunted house. Only audio was helpful
    - Recall the video of Family Guy, mentioning about “GILF”
2. **Performs better than most combinations, falling behind Image-modality alone**
    - Even so, performances are still acceptable, incorrect predictions are also “slight negatives”
3. **Problem statement was to protect children, logical to filter out more videos, even at an expense of sacrificing some videos**
    - This is however, mitigated by some of my work
        - E.g.: Voting Mechanism, where some negatives are “slight negative”, might opt to allow for such content

Limitations and Future Works

Limitations

1. **Scarcity of Dataset**
   - Need more data on “horror”, “Gore” etc
2. **Model capability limited to data trained on**
   - No capabilities of detecting pornographic materials visually
3. **Resource Intensive**
   1. Models of large sizes being used requires a lot of computational resources
   2. Might not be able to undergo one-shot-inference without sufficient resources

Future Works

1. **XAI**
   - For Audio mainly
2. **Comparing performance against MLLM**
   - Many multimodal LLMs released during this research (Eg: Video-LLaVA - end Jan 2024)
3. **Try larger models**
   - LoRa adapters finetuning
   - Explore Quantization

References

Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. Information Fusion, 76, 204–226.

Abi-Jaoude, E., Naylor, K. T., & Pignatiello, A. (2020). Smartphones, social media use and youth mental health. CMAJ, 192(6), E136–E141.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1), 1–6.

Cotter, K., DeCook, J. R., Kanthawala, S., & Foyle, K. (2022). In fyp we trust: The divine force of algorithmic conspirituality. International Journal of Communication, 16, 1–23.

CNA. (2023, November 9). https://www.channelnewsasia.com/business/uk-focuses-child-safety-start-new-online-regime-3908366?cid=internal_sharetool_androidphone_09112023_cna

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512–515).

Deininger, L., Stimpel, B., Yuce, A., Abbasi-Sureshjani, S., Schonenberger, S., Ocampo, P., ... & Gaire, F. (2022). A comparative study between vision transformers and CNNs in digital pathology. arXiv preprint arXiv:2206.00389.

Dosovitskiy, L., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

References

Foo, Y. C. (2023, November 9). *Reuters.com*. reuters.com. https://www.reuters.com/technology/cybersecurity/youtube-tiktok-be-asked-details-measures-protecting-minors-2023-11-08/

Gunawan, T. S., Ashraf, A., Riza, B. S., Haryanto, E. V., Rosnelly, R., Kartiwi, M., & Janin, Z. (2020). Development of video-based emotion recognition using deep learning with google colab. TELKOMNIKA (Telecommunication Computing Electronics and Control), 18(5), 2463–2471.

Kabali, H. K., Irigoyen, M. M., Nunez-Davis, R., Budacki, J. G., Mohanty, S. H., Leister, K. P., & Bonner Jr, R. L. (2015). Exposure and use of mobile media devices by young children. Pediatrics, 136(6), 1044–1050.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980–2988).

Morency, L.-P., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th international conference on multimodal interfaces (pp. 169–176).

Nguyen, H. T., Duong, P. H., & Cambria, E. (2019). Learning short-text semantic similarity with word embeddings and external knowledge sources. Knowledge-Based Systems, 182, 104842.

Ou, X., & Li, H. (2020). Ynu@ dravidian-codemix-fire2020: Xlm-roberta for multi-language sentiment analysis. In FIRE (Working Notes) (pp. 560–565).

References

Pant, K., & Dadu, T. (2020). Cross-lingual inductive transfer to detect offensive language. arXiv preprint arXiv:2007.03771.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., & Hussain, A. (2018). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. IEEE Intelligent Systems, 33(6), 17–25.

Razi, A. (2024, January 31). *Teens on social media need both protection and privacy – AI could help get the balance right*. The Conversation. https://theconversation.com/teens-on-social-media-need-both-protection-and-privacy-ai-could-help-get-the-balance-right-222052

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Ramesh, A. (2021). Learning transferable visual models from natural language supervision. In International conference on machine learning.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

Rideout, V., & Robb, M. B. (2017). The commonsense census: Media use by kids age zero to eight. San Francisco, CA: Common Sense Media, 263, 283.

Wang, J., Li, B., Hu, W., & Wu, O. (2011). Horror video scene recognition via multiple-instance learning. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1325–1328). IEEE.

Wu, T., Peng, J., Zhang, W., Zhang, H., Tan, S., Yi, F., ... & Huang, Y. (2022). Video sentiment analysis with bimodal information-augmented multihead attention. Knowledge-Based Systems, 235, 107676.

References

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1–5).

Yesilada, M., & Lewandowsky, S. (2022). Systematic review: Youtube recommendations and problematic content. Internet Policy Review, 11(1).

Zhang, M., Lucas, J., Ba, J., & Hinton, G. E. (2019). Lookahead optimizer: k steps forward, 1 step back. Advances in neural information processing systems, 32.

The END

THANK YOU